Workshop on Multi-PIM Systems for High-End Computing

Thomas Sterling and Sheila Vaidya
California Institute of Technology

Executive Summary

The need for trans-petaflop computing performance for applications ranging from national security to protein-folding and climate prediction, within the next 5-10 years, is widely acknowledged. An approach to its realization, while operating within practical constraints of size, power, cost, and reliability, must take a systems view where one simultaneously exploits advances in semiconductor and component technology, innovative architecture, and runtime and compiler software, to develop high-density systems of unprecedented performance. An enabling foundation for such a construct is an emerging class of integrated circuits referred to as processor-in-memory (PIM). PIM relies on the ability to fabricate both CMOS logic and DRAM (or SRAM) memory on the same die, permitting new configurations based on close physical and logical relationships between these two historically separate primary components. With a clever coupling of PIM with 3-d packaging and advanced cooling concepts, higher interconnect bandwidth, lower memory access latency, lower power dissipation, and smaller system size could result. From a national security standpoint, the challenge would then lie in selectively evolving those technologies which map over into the commercial arena, so as to ensure an industrial supplier for supercomputers commensurate with future progression of Moore's law.

With this goal in mind, a workshop on the implementation of multi-PIM systems was initiated by LLNL, to consider the opportunities, challenges, and strategies governing the application of PIM technology to a myriad of petaflop-scale applications ranging from radiation hydrodynamics and CFD to embedded, autonomous, space-borne, and large database manipulations. The objectives of the workshop were to

- Establish the potential value and opportunities afforded by PIM.
- Capture the state of the art, experience, and practices with PIM.
- Identify the critical challenges to the achievement of advanced PIM-based systems.
- Develop tenable strategies to fully exploit PIM.
- Devise a roadmap and recommendations for future development.

The Technical Committee for the Workshop consisted of Dave Cooper (LLNL), Bill Gropp (ANL), Peter Kogge (University.of Notre Dame), Bob Lucas (USC ISI), Jose Moreira (IBM), Burton Smith (Cray), Marc Snir (University of Illinois), Thomas Sterling (Caltech & JPL), Sheila Vaidya (LLNL), and Hans Zima(University of Vienna & JPL).

The workshop was sponsored by a number of institutions including Lawrence Livermore National Laboratory, National Security Agency, DOE-Defense Programs (ASCI), NASA, Sandia National Laboratories, Los Alamos National Laboratory, and the DOE Office of Science.

The three-day meeting (February 24-27, 2002), chaired by Dr. Thomas Sterling, involved a series of plenary presentations and short talks by leaders in the field of large-scale computing. Attendees constituted an interdisciplinary body of more than 50 experts from diverse backgrounds spanning microelectronics to systems engineering. Paul Messina of Caltech gave the opening keynote address, in which he emphasized the importance of

trans-petaflop-scale computing to national security. Burton Smith of Cray discussed system issues pertaining to the practical contributions of PIM to petaflops, including their impact on memory latency, bandwidth, temporal and spatial locality, and their interplay with ultra-lightweight threads and compilation techniques. Bill Gropp of ANL discussed features of programming models for PIM-based systems, while Marc Snir of the University of Illinois (ex-IBM) reminded us of the hurdles surrounding an industrial embrace of revolutionary architectures. From the federal end, Bob Graybill of DARPA (ITO) described the opportunities for evolving petaflop-scale computing in concert with his new High Productivity Computing Platform (HPCS) initiative, while Jose Muñoz and Fred Johnson covered the DOE ASCI and SciDac and Base Research programs. Working groups in the areas of Architecture and Systems (chaired by Peter Kogge), Applications and Programming Methods (chaired by Rick Stevens), and Runtime and Operating Systems (chaired by Dan Reed) then spent the bulk of the workshop analyzing the ramifications of PIM-based constructs in their respective areas.

The atmosphere was informal but the discussions were lively and the breakout groups worked well into the night. Monty Denneau of IBM discussed the petaflops opportunity based on the Blue Gene/Cyclops project; Christopher Krazyrakis and Kathy Yelick described the UC Berkeley IRAM project; Thomas Sterling described the design principles of the Gilgamesh MIND architecture, which incorporates fine grained multithreading and hardware supported "parcel- messaging," Joseph Torrellas of the University of Illinois presented the underlying concepts behind the FLEXRAM intelligent memory, and Bill Dally of Stanford talked about a PIM architecture for streaming supercomputing, that exploits both locality and concurrency for favorable performance to cost.

The Architecture Working Group contrasted PIM with SMP structures, combining communication bandwidth and latency into a "reachability" of access metric, from/to any given point in the system. System constructs included homogenous PIM arrays, PIMs used simply as local accelerators, PIMs as smart memory in a larger system with other computing resources, and PIMs as self centered agents creating aggregate logical structures through self organization with other PIMS without centralized control. The predominant questions with regards to implementing multi-PIM systems centered around packaging, cooling, interconnect topology, memory-processor ratio and configuration, internal and external communication bandwidth and latency, and the logical execution model. Memory capacity per PIM node, and consequently the number of nodes per PIM chip, was considered from both a coarse-grained and a fine-grained perspective. The coarse grained structure was advocated on the premise that a balance of bytes/flop was essential to limit off-chip communication. The fine-grained approach was supported to achieve as much on-chip memory bandwidth and thread parallelism as possible. The Group concluded that while PIMs offered significant opportunities, they would also impose major modifications to conventional practices in programming and resource management.

The Applications Team evaluated the applications space executable on PIM-based petaflop-scale systems and the methods and tools required to program them. Two different classes of programming models were considered: a high level, which would provide rapid application specification with description of forms of parallelism in both data and flow control, and a low-level language, which would give the programmer

performance transparency but control of resources, mapping, and access to network, memory, and processor scheduling.

The Runtime and Operating System Group started with a canonical architecture, with PIM chips and conventional compute processors internally connected into a shared memory infrastructure, and with multiple such units integrated by a global interconnect network through message passing. The software architecture encompassed the low level run time system, that operated local to a given PIM node, a global runtime system that operated across PIMs within a single unit, and the overall operating system that managed the combined resources, the external I/O environment, and system administrative responsibilities.

The degree of detail in these discussions not only highlighted the system potential of PIM, but also exposed the current dearth in understanding of PIM attributes and layout implications so as to devise meaningful roadmaps by which this technology could be incorporated into future petaflop-class systems.

The workshop concluded with the following set of recommendations.

- PIM holds sufficient promise for a practical petaflop-scale demonstration by the end of this decade. However, for effective insertion into the supercomputing backbone, with a credible industrial roadmap for delivery, federal agencies responsible for the advancement and deployment of high end computing should facilitate R&D on PIM-based architecture and supporting software infrastructure. (The DARPA ITO HPCS Program is a credible step in this direction.)
- A foundation study, which contrasts PIM-based architectures with more conventional ones and evolves the optimal one(s), is essential.
- Different classes of petaflop applications need to be classified on the basis of quantitative parameters such as memory capacity, interconnect bandwidth, data reuse locality, and program concurrency.
- Research into programming methodologies is necessary to establish a migration path for legacy codes to PIM-based systems and to devise advanced programming models that facilitate of PIM attributes.
- A parallel development of hybrid hardware/software-based fault-tolerant techniques which can ensure high availability of systems comprising millions of PIM nodes is essential.
- A new generation of runtime system software architecture that manages fine-grained PIM nodes with low spatial and temporal overhead must be pursued, and abstract interfaces to conventional operating system services and programming language compilers need to be devised.
- Experimental testbeds should be developed and distributed throughout the high end computing research community to expedite effective exploration of this field. Both software simulators and FPGA-based hardware platforms will be necessary, with a sharing of open source compilers, runtime software and tools, and results, in order to accelerate contributions within this new, enabling class of computer architecture.